Enhancing Image Watermarking Methods With/Without Reference Images by Optimization on Second-Order Statistics

Jengnan Tzeng, Wen-Liang Hwang, and I-Liang Chern

***Abstract*—The watermarking method has emerged as an important tool for content tracing, authentication, and data hiding in multimedia applications. We propose a watermarking strategy in which the watermark of a host is selected from the robust features of the estimated forged images of the host. The forged images are obtained from Monte Carlo simulations of potential pirate attacks on the host image. The solution of applying an optimization technique to the second-order statistics of the features of the forged images gives two orthogonal spaces. One of them characterizes most of the variations in the modifications of the host. Our watermark is embedded in the other space that most potential pirate attacks do not touch. Thus, the embedded watermark is robust. Our watermarking method uses the same framework for watermark detection with a reference and blind detection. We demonstrate the performance of our method under various levels of attacks.**

***Index Terms*—Authentication, copyright protection, watermark.**

I. Introduction

DIGITAL signatures embedded in contents, called "watermarks," are important for copyright protection, copyright control, and information hiding in multimedia applications [1], [10], [15], [22]. From the perspective of watermark detection, the media content where a watermark is embedded is noise. However, as pointed out in [8], the media content should not be viewed purely as noise since this view does not take advantage of the fact that the content is known completely to the watermark embedders. Thus, one should embed a watermark according to the available information of the content. This view of watermark embedment has a similar approach in signal selection for optimum coherent detection [20] where optimum signals to be embedded in a noisy channel whose properties are known to the sender, are selected. From this point of view, we demonstrate that there is a reasonable method for choosing an optimum watermark sequence according to the robust features of the content and the statistics of possible attacks. We propose a subspace watermarking method, where an optimum subspace of an image, from which a watermark sequence is selected for the image, is derived for detection according to the covariance of a Monte Carlo simulation of pirate attacks on the image.

Manuscript received June 28, 2000; revised February 27, 2002. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Naohisa Ohta.

J. Tzeng and W.-L. Hwang are with the Institute of Information Science, Academia Sinica, Taipei 115, Taiwan, R.O.C. (e-mail: whwang@iis.sinica.edu.tw).

I. L. Chern is with the Department of Mathematics, National Taiwan University, Taipei 115, Taiwan, R.O.C.

Publisher Item Identifier 10.1109/TIP.2002.800895.

Usually, watermarking methods are classified into two types: visible and invisible. We will focus our discussion on invisible watermarks. One can refer to [3] for a discussion of visible watermarks. Watermark *robustness* particularly refers to the ability to detect an embedded watermark in an image even when the image is modified by means of image operations. In [7], [18], and [19], there are many interesting discussions of watermark tampering methods. Despite much previous research, watermark robustness is still a worthwhile topic with plenty of unknown issues. Another important watermarking property, according to Craver *et al.* [9], is *ambiguity* regarding the retrieval of a watermark which is unambiguously identified by the owner. It has been shown that for a large class of watermarking methods that require the use of a reference image to identify ownership, there is ambiguity in resolving the rightful ownership of an image with multiple signatures [23]. This has motivated research on watermarking methods requiring no reference images in watermark detection process (blind detection) [24].

We propose a watermarking method which uses optimization methods to embed invisible watermarks in images. We assume that a pirate attack on an image aims to create an invisible modification of it by means of image operations. Like many previous researches, we embed our watermark using features. The features can be obtained from DCT coefficients, wavelet coefficients, spatial patterns (Dirac patterns), etc. We simulate pirate attacks by modifying our original image using image operations so that the resultant forged images are still visually acceptable. This means that our operations do not produce excessive visual quality loss. The probability density function of the features $\underline{\mathbf{e}}$ of the estimated forged images is obtainable. As a result, we can calculate the statistics of $\underline{\mathbf{e}}$ and, thus, characterize statistically the pirate attacks on the image. In practice, $\underline{\mathbf{e}}$ is a random variable in a fixed dimension vector space.

We then show that from the second-order statistics of $\underline{\mathbf{e}}$, we are able to partition the feature space where $\underline{\mathbf{e}}$ lies in two subspaces orthogonal to each other in such way that one of them, called V, has most of the variations of the estimated forged images. As a consequence, its complementary space, called W, contains fewer variations of $\underline{\mathbf{e}}$. Since $\underline{\mathbf{e}}$ characterizes our simulations of potential attacks, W will be the subspace in which attacks have less chance of modifying feature components. In other words, embedding a watermark within features in space W, called the watermark space, is more robust since doing so offers a higher probability of protecting the watermark against pirate attacks. Any vector in the watermark space can be our watermark feature of the original image. Our method can be used

in watermark detection no matter whether a reference image is required or not.

A word on the notations we use: a bold capital letter stands for a matrix (or an image), a bold small letter stands for a column vector, and an underlined letter denotes a random variable. Also, the transposition of a matrix $\mathbf{A}$ is $\mathbf{A}'$. We use $[\mathbf{A}]$ to denote arranging the matrix elements in $\mathbf{A}$ as a vector sequence in a pre-given order, for example, a lexicographic order. In Section II, we will specify our problem in a very general setting. In Section III-A, we will present our watermarking method which requires the use of a reference image for detection. The same framework will be extended in Section III-B to a watermark for blind detection. In Section IV, we will demonstrate the performance of our method subject to different levels of attack. Finally, a conclusion will be given in the last section.

II. Problem Model

If we are able to measure the effect of possible attacks on an image and embed our watermark in the features of the image that are resistant to the attacks, then the performance of watermark detection can be improved. Given a host image $\mathbf{X}$, of size $N = m \times n$ pixels, one can experiment on $\mathbf{X}$ by performing elementary image processing operations, such as translation, rotation, smoothing, compression, etc., or by combining elementary operations to obtain an estimated forged image of $\mathbf{X}$. We use the random variable $\underline{\mathbf{X}}$ for images obtained by means of such modifications of $\mathbf{X}$.

Let us expand $\underline{\mathbf{X}}$ against the bases $\{\mathbf{\Phi}_{i,j} | i = 1, \ldots, m, j = 1, \ldots, n\}$, whose dual bases are $\{\tilde{\mathbf{\Phi}}_{i,j} | i = 1, \ldots, m, j = 1, \ldots, n\}$. We have

$$\underline{\mathbf{X}} = \sum_{i,j} \langle \underline{\mathbf{X}}, \mathbf{\Phi}_{i,j} \rangle \tilde{\mathbf{\Phi}}_{i,j} \tag{1}$$

where $\langle \underline{\mathbf{X}}, \mathbf{\Phi}_{i,j} \rangle$ is the inner product of $\underline{\mathbf{X}}$ and the basis function $\mathbf{\Phi}_{i,j}$. Equation (1) can be written as

$$\underline{\mathbf{X}} = \sum_{i,j} \langle \underline{\mathbf{X}} - \mathbf{X}, \mathbf{\Phi}_{i,j} \rangle \tilde{\mathbf{\Phi}}_{i,j} + \sum_{i,j} \langle \mathbf{X}, \mathbf{\Phi}_{i,j} \rangle \tilde{\mathbf{\Phi}}_{i,j}$$

where $\langle \underline{\mathbf{X}} - \mathbf{X}, \mathbf{\Phi}_{i,j} \rangle$ is the deviation of $\underline{\mathbf{X}}$ from $\mathbf{X}$ along the basis $\mathbf{\Phi}_{i,j}$. The corresponding feature is denoted as

$$\underline{\mathbf{e}} = [\langle \underline{\mathbf{X}} - \mathbf{X}, \mathbf{\Phi}_{i,j} \rangle]. \tag{2}$$

Let $\mathbf{X}^{\mathbf{M}} = \mathbf{X} + \mathbf{M}$ be the watermarked image obtained by embedding watermark $\mathbf{M}$ into image $\mathbf{X}$. As before, we use $\underline{\mathbf{X}}^{\mathbf{M}}$ to denote the modifications of $\mathbf{X}^{\mathbf{M}}$ by means of elementary image operations and their combinations

$$\begin{aligned} \underline{\mathbf{X}}^{\mathbf{M}} &= \sum_{i,j} \langle \underline{\mathbf{X}}^{\mathbf{M}}, \mathbf{\Phi}_{i,j} \rangle \tilde{\mathbf{\Phi}}_{i,j} \\ &= \sum_{i,j} \langle \mathbf{X}, \mathbf{\Phi}_{i,j} \rangle \tilde{\mathbf{\Phi}}_{i,j} \\ &\quad + \sum_{i,j} \left(\langle \mathbf{X}^{\mathbf{M}} - \mathbf{X}, \mathbf{\Phi}_{i,j} \rangle + \langle \underline{\mathbf{X}}^{\mathbf{M}} - \mathbf{X}^{\mathbf{M}}, \mathbf{\Phi}_{i,j} \rangle \right) \tilde{\mathbf{\Phi}}_{i,j}. \end{aligned}$$

Every item on the right side of the above equation is explained: If we use inner product coefficients as our features, then $[\langle \mathbf{X}, \mathbf{\Phi}_{i,j} \rangle]$ represents the features of the host image, and the vector

$$\mathbf{m} = [\langle \mathbf{X}^{\mathbf{M}} - \mathbf{X}, \mathbf{\Phi}_{i,j} \rangle] \tag{3}$$

is the watermark features added by copyright owners. We are able to obtain the watermark $\mathbf{M}$ from $\mathbf{m}$ by means of

$$\mathbf{M} = \sum_{i,j} \langle \mathbf{X}^{\mathbf{M}} - \mathbf{X}, \mathbf{\Phi}_{i,j} \rangle \tilde{\mathbf{\Phi}}_{i,j}. \tag{4}$$

The last term on the right side, $[\langle \underline{\mathbf{X}}^{\mathbf{M}} - \mathbf{X}^{\mathbf{M}}, \mathbf{\Phi}_{i,j} \rangle]$, has a special meaning; it represents the features which deviate from those of $\mathbf{X}^{\mathbf{M}}$ and possibly were introduced by a pirate attack. We denote the feature perturbation from that of $\mathbf{X}^{\mathbf{M}}$ as

$$\underline{\mathbf{e}}^{M} = [\langle \underline{\mathbf{X}}^{\mathbf{M}} - \mathbf{X}^{\mathbf{M}}, \mathbf{\Phi}_{i,j} \rangle] \tag{5}$$

and the centered perturbation as

$$\underline{\mathbf{e_c}}^{M} = [\langle \underline{\mathbf{X}}^{\mathbf{M}} - \mathbf{X}^{\mathbf{M}}, \mathbf{\Phi}_{i,j} \rangle] - [\langle E\{\underline{\mathbf{X}}^{\mathbf{M}}\} - \mathbf{X}^{\mathbf{M}}, \mathbf{\Phi}_{i,j} \rangle]. \tag{6}$$

We are interested in the following watermark selection problem: Which watermark feature $\mathbf{m}$ to be embedded in the host is most resistant to the random feature perturbation, $\underline{\mathbf{e}}^{M}$, introduced by possible attacks on the host image?

A. Remark

In many researches, a subset S of $\{\mathbf{\Phi}_{i,j}\}$ with coefficients more relevant to perceptual substances or more robust to statistical decision are selected, where $\{\mathbf{\Phi}_{i,j}\}$ can be either DCT bases [6], wavelet bases [13], or delta functions if the watermark is to be embedded by means of spatial domain methods [2], [21]. Copyright owners can then embed their watermarks only into the coefficients in S (its complement set is $\bar{S}$); therefore

$$\mathbf{X}^{\mathbf{M}} = \sum_{\mathbf{\Phi}_{i,j} \in S} \langle \mathbf{X}^{\mathbf{M}}, \mathbf{\Phi}_{i,j} \rangle \tilde{\mathbf{\Phi}}_{i,j} + \sum_{\mathbf{\Phi}_{i,j} \in \bar{S}} \langle \mathbf{X}, \mathbf{\Phi}_{i,j} \rangle \tilde{\mathbf{\Phi}}_{i,j}.$$

Since watermark information is contained entirely in the coefficients of S, the features we need for embedding a watermark are restricted to S. The restricted watermark feature is denoted as $\mathbf{m}|_S = [\langle \mathbf{X}^{\mathbf{M}} - \mathbf{X}, \mathbf{\Phi}_{i,j} \rangle]_{\mathbf{\Phi}_{i,j} \in S}$. The to-be-proposed watermarking methods are applicable to features extracted either from all bases $\{\mathbf{\Phi}_{i,j}\}$ or from any subset S. Thus, for simplicity, in most of the following sections, we will assume that watermarks are cast on features relevant to all bases $\{\mathbf{\Phi}_{i,j}\}$. Only in Section IV, which will cover our experiments, will we use a subset S of the DCT bases.

III. Watermark Subspace Selection

Many measurements have been proposed for watermark detection [16]. Among them, a frequently used one is the correlation measurement, which measures the cosine angle of the two feature vectors, $\mathbf{u}$ and $\mathbf{v}$, by means of

$$\mathrm{sim}(\mathbf{u}, \mathbf{v}) = \frac{|\mathbf{u}'\mathbf{v}|}{||\mathbf{u}||||\mathbf{v}||}.$$

We say that the two vectors are similar if their sim value is close to one; or we say that the two are not similar if this value is close

to zero. Let $\mathbf{X}$ be the host image. We aim to select the watermark feature $\mathbf{m}$ of $\mathbf{X}$ and a linear transform $\mathbf{P}$, which is related to $\mathbf{m}$, such that we have the following.

High detection probability: If the feature $\mathbf{t}$ is extracted from an attacked watermarked image of $\mathbf{X}$, then the $\mathrm{sim}(\mathbf{m}, \mathbf{P}'\mathbf{t})$ should be as large as possible.

Low false-alarm probability: If the feature $\mathbf{t}$ is extracted from an unwatermarked image, then the $\mathrm{sim}(\mathbf{m}, \mathbf{P}'\mathbf{t})$ should be as small as possible.

We propose a watermarking strategy in which $\mathbf{m}$ is embedded in a subspace W, called the watermark subspace, which is robust against the pirate attacks, and where $\mathbf{P}$ is a projection onto this subspace. To find the watermark subspace in which we embed our watermark feature such that the conditions of high detection probability and low false-alarm probability are met in watermark detection, some optimization method should be applied to an objective function related to our watermark detection scheme.

A. Watermark Subspace for Detection With a Reference

Suppose our feature space is R^N, and that our watermark feature $\mathbf{m} \in W$. The feature perturbation $\underline{\mathbf{e}}^M$, after projecting to subspace W, can be rewritten as

$$P_W(\underline{\mathbf{e}}^M) = \underline{\alpha}\mathbf{m} + \underline{\mathbf{w}}$$

where $\underline{\alpha}$ is a scalar random variable, obtained by projecting $P_W(\underline{\mathbf{e}}^M)$ onto $\mathbf{m}$, and $\mathbf{m}$ and $\underline{\mathbf{v}}$ are perpendicular to each other. If W is the subspace of R^N such that most of the realizations of $\underline{\mathbf{v}}$ have

$$\begin{cases} (a) & \|\underline{\mathbf{w}}\| \ll \|\mathbf{m}\| \\ (b) & |\underline{\alpha}| \text{ is close to } 0 \end{cases} \tag{7}$$

then for most $\underline{\mathbf{w}}$, we will have the following detection probability and false alarm probability.

Detection Probability: If a test image contains our watermark sequence $\mathbf{m}$, and if our watermark detection method requires a reference image, then the correlation measurement between $\mathbf{m}$ and the projection of an extracted feature $\underline{\mathbf{t}}$ onto W is

$$\begin{aligned} \mathrm{sim}(\mathbf{m}, P_W(\underline{\mathbf{t}})) &= \mathrm{sim}\left(\mathbf{m}, P_W(\mathbf{m} + \underline{\mathbf{e}}^M)\right) \\ &= \frac{\left|\mathbf{m}' P_W(\mathbf{m} + \underline{\mathbf{e}}^M)\right|}{\|\mathbf{m}\| \|P_W(\mathbf{m} + \underline{\mathbf{e}}^M)\|}. \end{aligned} \tag{8}$$

According to (7a), we have

$$\begin{aligned} \mathrm{sim}(\mathbf{m}, P_W(\mathbf{m} + \underline{\mathbf{e}}^M)) &= \mathrm{sim}(\mathbf{m}, (1+\underline{\alpha})\mathbf{m} + \underline{\mathbf{w}}) \\ &\approx \mathrm{sim}(\mathbf{m}, (1+\underline{\alpha})\mathbf{m}) = 1. \end{aligned}$$

Hence, if $\underline{\mathbf{e}}^M$ represents the features corresponding to attacks on $\mathbf{X}^{\mathbf{M}}$, and if $\mathbf{m}$ is chosen so as to satisfy (7a), then the watermark detection probability is high.

False Alarm Probability: If a test image does not contain $\mathbf{m}$, and if we assume that an attack on the unwatermarked test will yield a result in W similar to which would be obtained if the attack were applied to the watermarked image, then the correlation measurement will be

$$\mathrm{sim}(\mathbf{m}, P_W(\underline{\mathbf{t}})) \approx \mathrm{sim}\left(\mathbf{m}, P_W(\underline{\mathbf{e}}^M)\right) = \frac{\left|\mathbf{m}' P_W(\underline{\mathbf{e}}^M)\right|}{\|\mathbf{m}\| \|P_W(\underline{\mathbf{e}}^M)\|}. \tag{9}$$

According to (7b), we have

$$\mathrm{sim}(\mathbf{m}, P_W(\underline{\mathbf{e}}^M)) = \frac{\alpha\|\mathbf{m}\|^2}{\alpha\|\mathbf{m}\|^2 + \|\mathbf{m}\|\|\mathbf{w}\|} \approx 0.$$

Thus, the false alarm probability is small. If our test image is the host image $\mathbf{X}$, then $\underline{\mathbf{t}}$ is $\underline{\mathbf{e}}$ which is given in (2).

1) Subspace Selection by Means of Second-Order Statistics: The conditions in (7) can be satisfied if W is chosen perpendicular to most of the realizations of $\underline{\mathbf{e}}^M$. We will resort to the second order statistics of $\underline{\mathbf{e}}^M$ to find the optimal watermark subspace W by means of the following objective function:

$$\min_{\mathbf{m} \in \mathbf{W}} E\{(\mathbf{m}' \underline{\mathbf{e_c}}^M)(\mathbf{m}' \underline{\mathbf{e_c}}^M)'\} \tag{10}$$

where $\underline{\mathbf{e_c}}^M$ is the centered perturbation of $\mathbf{X}^{\mathbf{M}}$ given in (6). By means of simple calculation, we have

$$\begin{aligned} E\{(\mathbf{m}' \underline{\mathbf{e_c}}^M)(\mathbf{m}' \underline{\mathbf{e_c}}^M)'\} &= \mathbf{m}' E\{(\underline{\mathbf{e_c}}^M)(\underline{\mathbf{e_c}}^M)'\}\mathbf{m} \\ &= \mathbf{m}'\mathbf{U}\mathbf{\Sigma}\mathbf{U}'\mathbf{m} \\ &= \sum_{i=1}^{N} \sigma_i^2 (\mathbf{m}'\mathbf{u}_i)^2 \end{aligned} \tag{11}$$

where $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1^2, \sigma_2^2 \ldots, \sigma_N^2)$, in which the eigenvalues σ_i^2 are arranged in decreasing order of magnitude and $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N]$ is an orthonormal matrix containing the principle components of the covariance matrix of $\underline{\mathbf{e_c}}^M$. If we separate the eigenvalues into nonzero and zero components

$$\begin{aligned} NZ &= \{\sigma_i^2 | \sigma_i^2 > 0; i = 1, \ldots, d\}, \\ Z &= \{\sigma_i^2 | \sigma_i^2 = 0; i = d+1, \ldots, N\} \end{aligned}$$

and if Z is not an empty set, then we can obtain a simple optimization of (10) by assigning our watermark feature $\mathbf{m}$ according to

$$\mathbf{m}'\mathbf{u}_k = \begin{cases} 0, & \text{if } \sigma_k^2 \in NZ \\ \text{arbitrary number}, & \text{if } \sigma_k^2 \in Z \end{cases}$$

and $\sum_k (\mathbf{m}'\mathbf{u}_k)^2 = \|\mathbf{m}\|^2 = C$, where C is a parameter relevant to the perceptual capacity of image $\mathbf{X}$. Usually, C is chosen so as to be large enough, but so that the resultant watermark is still invisible. Any vector $\mathbf{m}$ in the subspace spanned by the orthonormal bases $\{\mathbf{u}_i | i = d+1, \ldots, N\}$ is a solution of (10), and it can be chosen as the watermark feature for $\mathbf{X}$. Thus, the optimal watermark space W, where the watermark feature $\mathbf{m}$ lies, is spanned by the orthonormal bases $\{\mathbf{u}_i | i = d+1, \ldots, N\}$ with dimension $N - d$.

For the case where there is no zero eigenvalue, one can obtain a solution of (10) according to the theorem of arithmetic and

geometric means. Assume that $\mathbf{m}'\mathbf{u}_i \neq 0$ for all i; we have the following inequality:

$$\frac{\sum_i \sigma_i^2(\mathbf{m}'\mathbf{u}_i)^2}{N} \geq \left(\prod_i \sigma_i^2(\mathbf{m}'\mathbf{u}_i)^2\right)^{1/N}. \quad (12)$$

This inequality becomes an equality if and only if

$$\frac{\sigma_i^2}{\sum_{l=1}^{N}\sigma_l^2}(\mathbf{m}'\mathbf{u}_i)^2 = \frac{\sigma_j^2}{\sum_{l=1}^{N}\sigma_l^2}(\mathbf{m}'\mathbf{u}_j)^2, \text{ for any pair of } \sigma_i^2 \text{ and } \sigma_j^2.$$

Thus,

$$(\mathbf{m}'\mathbf{u}_i)^2 = \frac{C}{\sigma_i^2\left(\sum_{l=1}^{N}\frac{1}{\sigma_l^2}\right)}. \quad (13)$$

We have

$$\mathbf{m}'\mathbf{u}_i = \pm\frac{1}{\sigma_i}\sqrt{\frac{C}{\left(\sum_{l=1}^{N}\frac{1}{\sigma_l^2}\right)}}$$

for $i = 1 \ldots N$ for any given C. The strength of $\mathbf{m}$ in the direction of a given eigenvector will be inversely proportional to the corresponding eigenvalue. There are 2^N choices for our watermark feature. Our watermark space is, thus, a finite set, whose dimension is zero.

Fixed-Dimension Watermark Subspace: The dimension of our watermark space should not be too small since malicious attackers can jam the watermark space by spreading random noise into it. For this reason, the dimension of the watermark space should be large enough to avoid such attacks. In practice, it is convenient to fix the dimension of W, say D, and to choose W such that it is spanned by the eigenvectors corresponding to the D smallest eigenvalues in $\{\sigma_k^2 | k = 1, \ldots, N\}$. This corresponds to finding a linear transformation of $\underline{\mathbf{e_c}}^M$ with a matrix $\mathbf{A}$ as

$$\mathbf{A}'\underline{\mathbf{e_c}}^M$$

where $\mathbf{A}$ is an N by D matrix, whose rank is D with $D \leq N$ and where each column of $\mathbf{A}$ has only one nonzero element with a value of one. The covariance matrix of the resultant random vector is

$$E\{\mathbf{A}'\underline{\mathbf{e_c}}^M\underline{\mathbf{e_c}}^{M'}\mathbf{A}\}$$

which is equal to $\mathbf{A}'\mathbf{U}\mathbf{\Sigma}\mathbf{U}'\mathbf{A}$. The sum of the D smallest eigenvalues in $\mathbf{U}\mathbf{\Sigma}\mathbf{U}'$ corresponds to finding the $\mathbf{A}$ such that the following objective function is minimized:

$$\min_{\mathbf{A}} \text{trace}(\mathbf{A}'\mathbf{U}\mathbf{\Sigma}\mathbf{U}'\mathbf{A})$$

where *trace* is the trace operation on a matrix. One can easily see that this corresponds to our optimum solution of (10) if we set the smallest D eigenvalues to zero and use the corresponding eigenvectors as the bases for our watermark space. Our watermark space W is characterized by the eigenvalues of the covariance of $\underline{\mathbf{e}}^M$. If the variations are concentrated mostly in the subspace V spanned by a few eigenvectors, whose corresponding eigenvalues take a large proportion of the total variance of $\underline{\mathbf{e}}^M$, then we expect that the image distortions due to pirate attacks will tend to produce images with features in subspace V. Thus, by superimposing a watermark with features mainly in the complementary subspace of V, called W, we can obtain a robust watermark in the sense that this watermark will have less chance of being erased by pirate attacks.

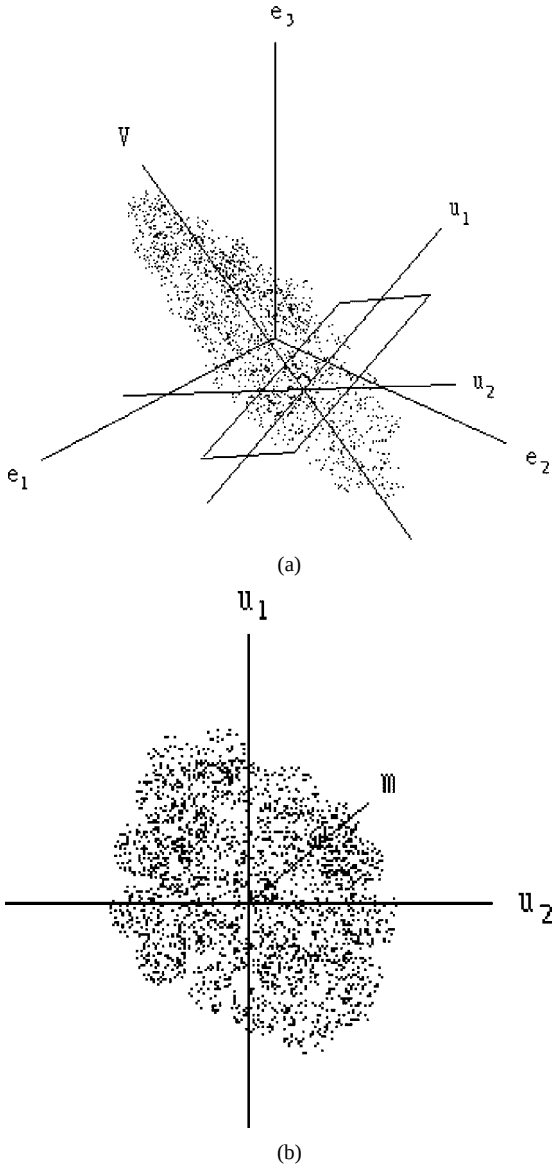


Fig. 1. Simplified schematic diagram of our watermarking strategy.

Fig. 1 is a simple schematic presentation of a watermark space with dimension 2 in R^3. From Fig. 1(a), we can see that most of the variation of $\underline{\mathbf{e}}^M$ lies in the subspace V, which has dimension 1. Its complementary space W has dimension 2, as shown in Fig. 1(b). $\mathbf{u}_1$ and $\mathbf{u}_2$ are the bases of W. Any vector that belongs to the watermark space W can be chosen as our watermark feature.

2) Perceptual Model: We will present a simple extension of the described method using a perceptual model. Previously, we stated that $[\mathbf{A}]$ is a sequence whose elements come from the matrix $\mathbf{A}$ in a pre-assigned order. Here, we use l to denote this

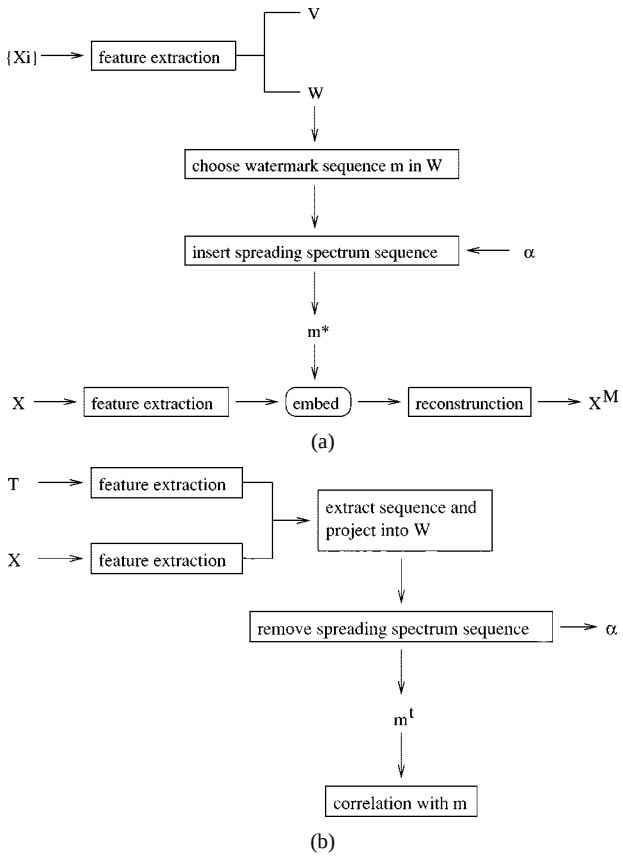


Fig. 2. (a) Our watermark encoder requires a reference image $\mathbf{X}$, where $\{\mathbf{X}_i\}$ stands for our estimated forged images. Our watermark feature, modulated using the spreading spectrum technique, is embedded in W. (b) Our watermark decoder requiring a reference image. $\mathbf{m}^t$ is obtained after the spreading spectrum random sequence is removed.

order. We say that the element at the location $l(i,j)$ in vector $[\mathbf{A}]$, denoted as $[\mathbf{A}]_{l(i,j)}$, is the element in the ith row and the jth column of $\mathbf{A}$, i.e., $\mathbf{A}_{i,j} = [\mathbf{A}]_{l(i,j)}$.

If a perceptual model is considered, we can write the watermarked coefficient at the basis $\mathbf{\Phi}_{i,j}$

$$\begin{aligned}\langle \mathbf{X}^{\mathbf{M}}, \mathbf{\Phi}_{i,j}\rangle &= \langle \mathbf{X}, \mathbf{\Phi}_{i,j}\rangle + \langle \mathbf{X}^{\mathbf{M}} - \mathbf{X}, \mathbf{\Phi}_{i,j}\rangle \\ &= \langle \mathbf{X}, \mathbf{\Phi}_{i,j}\rangle + \mathbf{m}_{l(i,j)} \\ &= \langle \mathbf{X}, \mathbf{\Phi}_{i,j}\rangle + r(\mathbf{\Phi}_{i,j})\mathbf{w}_{l(i,j)} \end{aligned} \tag{14}$$

where $\mathbf{m}_{l(i,j)}$ and $\mathbf{w}_{l(i,j)}$ are the $l(i,j)$th element of the vectors $\mathbf{m}$ and $\mathbf{w}$, respectively, and $\mathbf{w}$ is a sequence whose value is between -1 and 1 in each component. Also, $r(\mathbf{\Phi}_{i,j})$ is a scalar indicating the sensitivity of the basis $\mathbf{\Phi}_{i,j}$ to the human visual system (HVS), whose value depends on which perceptual model and which bases have been used. In a perceptual model, $|\mathbf{m}_{l(i,j)}|$ is constrained to be no larger than the just-noticeable-distortion (JND) value of the basis $\mathbf{\Phi}_{i,j}$. If we choose $r(\mathbf{\Phi}_{i,j})$ to be the JND of the basis, then $\mathbf{w}_{l(i,j)}$ will be a sequence whose value is either -1 or 1 in each component. The reader may refer to [4] and [11] for a discussion of JND in Fourier bases, in wavelet bases, and in spatial domain patterns.

3) Watermark Encoding: Fig. 2(a) shows our watermark encoding method. Any selected vector $\mathbf{m} \in W$ can be either a random sequence or a logo. The spreading spectrum technique is then applied to modulate $\mathbf{m}$ with a random sequence. Let $\mathbf{m}^* \in W$ be the resultant watermark feature. Since $\mathbf{X}^{\mathbf{M}^*} = \mathbf{X} + \mathbf{M}^*$, according to (3), we have

$$\mathbf{m}^* = \left[\langle \mathbf{X}^{\mathbf{M}^*} - \mathbf{X}, \mathbf{\Phi}_{i,j}\rangle\right] = P_W\left(\left[\langle \mathbf{X}^{\mathbf{M}^*} - \mathbf{X}, \mathbf{\Phi}_{i,j}\rangle\right]\right). \tag{15}$$

In Section IV, an example is shown where a subset of DCT coefficients is used as our feature; therefore, both $\mathbf{\Phi}$ and $\tilde{\mathbf{\Phi}}$ are DCT bases. The unitary matrix $\mathbf{U}$ is obtained from the covariance of the features of the estimated forged images of $\mathbf{X}$. A fixed-dimension watermark subspace W is then selected from this $\mathbf{U}$ according to the magnitudes of its eigenvalues.

4) Watermark Decoding: Given a test image $\mathbf{T}$, we first subtract the reference image, which is $\mathbf{X}$, from the test image and then represent the resultant image using the bases $\{\mathbf{\Phi}_{i,j}\}$ (in Section IV, we will use DCT bases). The extracted feature $\mathbf{t}$ is

$$\begin{aligned}\mathbf{t} &= [\langle \mathbf{T} - \mathbf{X}, \mathbf{\Phi}_{i,j}\rangle] \\ &= [\langle \mathbf{X}^{M^*} - \mathbf{X}, \mathbf{\Phi}_{i,j}\rangle] + [\langle \mathbf{T} - \mathbf{X}^{M^*}, \mathbf{\Phi}_{i,j}\rangle].\end{aligned}$$

Projecting $\mathbf{t}$ onto the space W, we have

$$P_W(\mathbf{t}) = \mathbf{m}^* + P_W([\langle \mathbf{T} - \mathbf{X}^{\mathbf{M}^*}, \mathbf{\Phi}_{i,j}\rangle]).$$

We then test $\mathrm{sim}(\mathbf{m}^*, P_W(\mathbf{t}))$ against a threshold and claim the ownership of the test image $\mathbf{T}$ when this value is greater than the threshold. For decoding, in addition to the host image $\mathbf{X}$, we require knowledge of the bases of the watermark space W and the sequence $\mathbf{m}^*$. A schematic diagram of our watermark decoding method requiring a reference image is shown in the bottom subfigure of Fig. 2. Here, we use $\mathbf{m}^*$ for watermark sequence $\mathbf{m}$ after conducting spreading spectrum modulation. The correlation drawn in this figure is carried out after the spreading spectrum random sequence is removed, which yields the same result as described previously.

B. Watermark Subspace for Blind Detection

Our watermarking method discussed previously can be modified for blind detection. If the feature $\underline{\mathbf{t}}$ is extracted from an attacked watermarked image, then

$$\mathrm{sim}(\mathbf{m}^*, P_W(\underline{\mathbf{t}})) = \mathrm{sim}\left(\mathbf{m}^*, P_W(\mathbf{m}^* + [\langle \mathbf{X}, \mathbf{\Phi}_{i,j}\rangle] + \underline{\mathbf{e}}^M)\right) \tag{16}$$

else

$$\mathrm{sim}(\mathbf{m}^*, P_W(\underline{\mathbf{t}})) = \mathrm{sim}\left(\mathbf{m}^*, P_W([\langle \mathbf{X}, \mathbf{\Phi}_{i,j}\rangle] + \underline{\mathbf{e}}^M)\right). \tag{17}$$

The arguments of the sim function in the above equations having a host image component are that, unlike the previous watermark decoder [see (8)], $[\langle \mathbf{X}, \mathbf{\Phi}_{i,j}\rangle]$ is unknown to the blind watermark decoder. Thus, the host component must be included in watermark detection.

If we choose W such that

$$P_W([\langle \mathbf{X}, \mathbf{\Phi}_{i,j}\rangle]) = 0$$

then (16) and (17) are reduced to (8) and (9), respectively. Thus, we use the method described in Section III-A3 to find a space W' with a given dimension $D > 2$. Then, we can find our

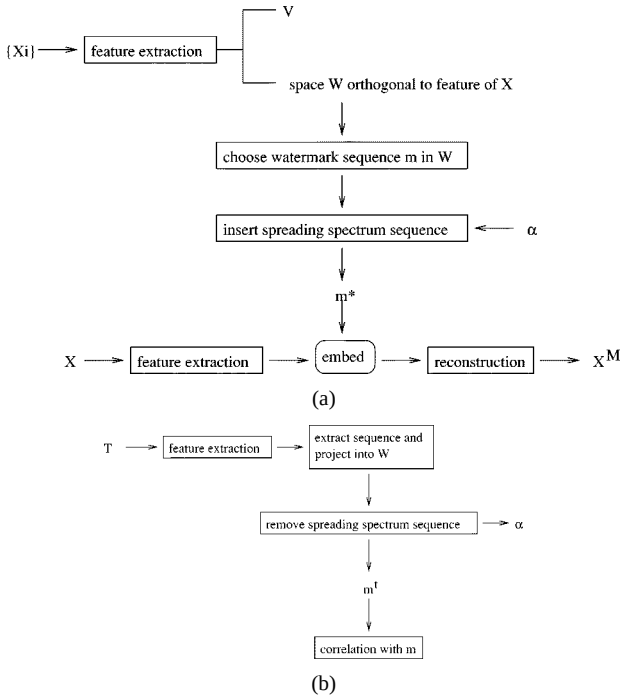


Fig. 3. (a) Our blind watermark encoder. The watermark sequence is embedded in the space W which is orthogonal to the feature of the host image X. (b) Our blind watermark decoder.

watermark space W as a subspace of W' such that W will also be orthogonal to $[\langle \mathbf{X}, \mathbf{\Phi}_{i,j} \rangle]$.

In our implementation, we first project $[\langle \mathbf{X}, \mathbf{\Phi}_{i,j} \rangle]$ onto W' and then choose the projected vector $P_{W'}([\langle \mathbf{X}, \mathbf{\Phi}_{i,j} \rangle])$ as the first vector to which we will apply Gram–Schmidt in W'. The bases in W' are again modified by Gram–Schmidt, which results in new bases orthogonal to the projected vector. These new bases are the bases of our watermark space W. If there is no basis in W' parallel to $P_{W'}([\langle \mathbf{X}, \mathbf{\Phi}_{i,j} \rangle])$, then the dimension of W is one less than that of W'; otherwise, the dimension of W' will be reduced by 2. We can derive from $W' \supset W$ and $P_W(P_{W'}([\langle \mathbf{X}, \mathbf{\Phi}_{i,j} \rangle])) = 0$ that

$$P_W([\langle \mathbf{X}, \mathbf{\Phi}_{i,j} \rangle]) = 0.$$

Given a threshold, the ownership of the test image $\mathbf{T}$ can be determined by comparing $\mathrm{sim}(\mathbf{m}^*, P_W(\mathbf{t}))$ to the threshold, where $\mathbf{t} = [\langle \mathbf{T}, \mathbf{\Phi}_{i,j} \rangle]$. Shown in Fig. 3 are block diagrams of our watermark encoding and decoding methods for blind detection, respectively. In decoding, the bases of the watermark space W and $\mathbf{m}^*$ are required.

IV. Experimental Results

In this section, we will demonstrate the attack resistance of our watermarking methods. The reader can refer to [14] and [19] for further discussions of various attacks. We applied full frame DCT to a set of 22 images including the images Lena, Barbara, and Mandrill. The major features of the other images are flowers, cats, dogs, forest, boys, and girls. Their sizes are all larger than 32×32. We then selected DCT coefficients from their upper left 32×32 corner, corresponding to combinations of 32 horizontal low-frequency bands and 32 vertical low-frequency bands. Thus, our vector space had a dimension of 1024. Then, we operated on each image to obtain a set of forged images by means of image operations. Our operations included blurring, compression with JPEG, small rotations (by $\pm 0.1^\circ$, $\pm 0.2^\circ$), small translations (by shifting one pixel either up, down, left or right), applying geometrical deformation (see http://www.cl.cam.ac.uk/ mgk25/stirmark/) to the image, adding random noise, and other image operations built into the image toolbox of Matlab and Microsoft Photo Editor. In total, we obtained 183 forged images for each image.

For each image, we then computed the covariance matrix from the collections of features obtained from the forged images of the image. Using singular value decomposition (SVD), we chose our watermark spaces with a fixed-dimension of 900 for each image. Each image has two watermark spaces: one for detection with a reference and the other for blind detection. The SVD results show that most of the eigenvalues were zero since the number of training feature vectors, which was 183 for each image was less than the dimension of a feature. We then studied the performance of our watermarking scheme by either embedding a visually meaningful pattern or by using a random sequence in the watermark space W of each image. Our pattern is a 30×30 Bee pattern [see Fig. 6(b)] taking a value in either 1 or -1. Embedding a meaningful pattern into space W provides more meaningful evidence to a judge than a detected number for verifying ownership.

A. Enhancing Existing Watermarking Methods

Our method can be used to enhance the strength of many existing watermarking methods. As an example, we will improve the robustness of the frequency domain watermarking method proposed by Cox *et al.* [6]. The method of Cox *et al.* uses DCT coefficients as features and requires a reference image for watermark detection. Their watermark feature is embedded in the significant DCT coefficients in a multiplicative way. Let v_i be the ith feature of the host image. Cox's method modifies v_i and obtains, for each i

$$v_i(1 + \alpha_i w_i).$$

This method can be implemented by adding to each feature component i a noise n_i with $n_i = v_i \alpha_i w_i$, thus obtaining

$$v_i + n_i$$

where the feature $\mathbf{n} = [n_1, n_2, \ldots, n_D]' \in R^N$. When our method is used, the resultant ith component is

$$v_i + m_i^*$$

where D is the dimension of watermark space W and our watermark feature $\mathbf{m}^* = [m_1^*, m_2^*, \ldots, m_D^*]'$ is in a subspace W of R^N.

In Fig. 4 are shown the mean and the standard deviation of the sim values obtained by performing various attacks on each image in our image set using our watermarking methods and that of Cox *et al.*, respectively. In these experiments, we kept the watermark energy and the watermark dimension the same for both methods. Attacks 1 to 5 were operations whose parameters were included in our obtained forged images, while the parameters of the operations used in Attacks 6 to 10 were larger

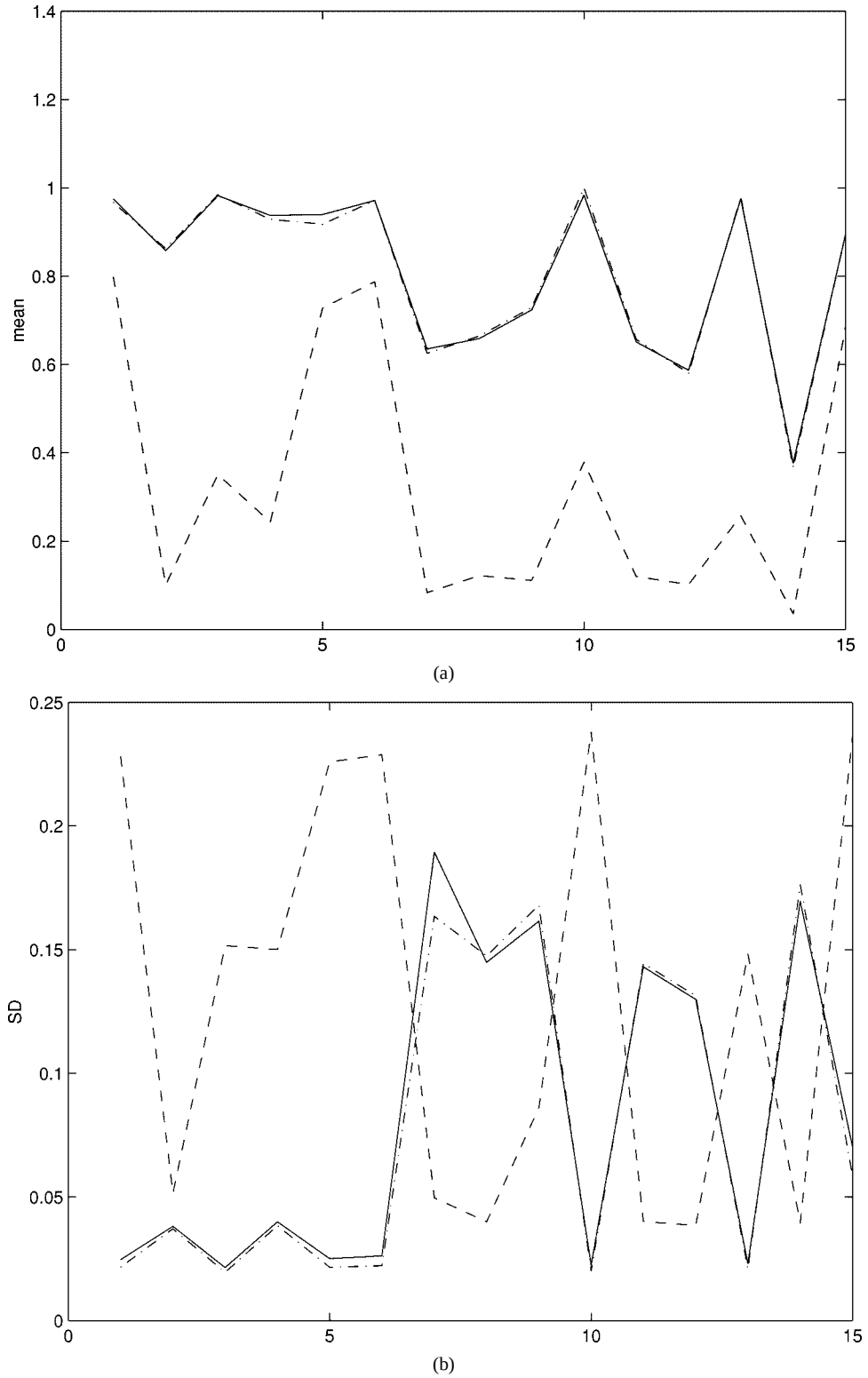


Fig. 4. Comparisons of the mean and standard derivation of various attacks on our methods (solid line with reference, dash-dot lines without reference) and on Cox's method (dash lines) with 22 test images. Each image was subjected to 15 attacks. The first five were operations that were intended to obtain our watermark space W, while the middle five were not, and the last five were combinations of attacks with one of them from one to five except for Attack 13. Attacks 1 to 5 were, respectively: 1) JPEG (60%): JPEG compression with a quality setting of 60%; 2) Stirmark(with small values for its parameters); 3) Small rotation 0.02°; 4) Small translation (one pixel in either direction); 5) Small random noise. Attacks 6 to 10 were: 6) JPEG (53%): JPEG compression with a quality setting of 53%; 7) Stirmark(with larger values than Attack 2); 8) Rotation 1°; 9) Translation two pixels in either direction; 10) Blur (cubic): Smooth by cubic spline. The last five were, respectively: 11) JPEG 60% + Rotation 1°; 12) Translation one pixel + Blur (cubic); 13) Rotate 10° and then rotate 10° back + blur (quadratic); 14) Stirmark (with the same parameters used in Attack 2) + Translation (two pixels); 15) Random noise (more noise than in Attack 5) + JPEG 53%.

than those used for forged images. Attacks 11 to 15 were combinations of various attacks. One can see from the mean and standard deviation, shown, respectively, in Fig. 4(a) and (b), the robustness of these methods to various attacks. Our methods, in-

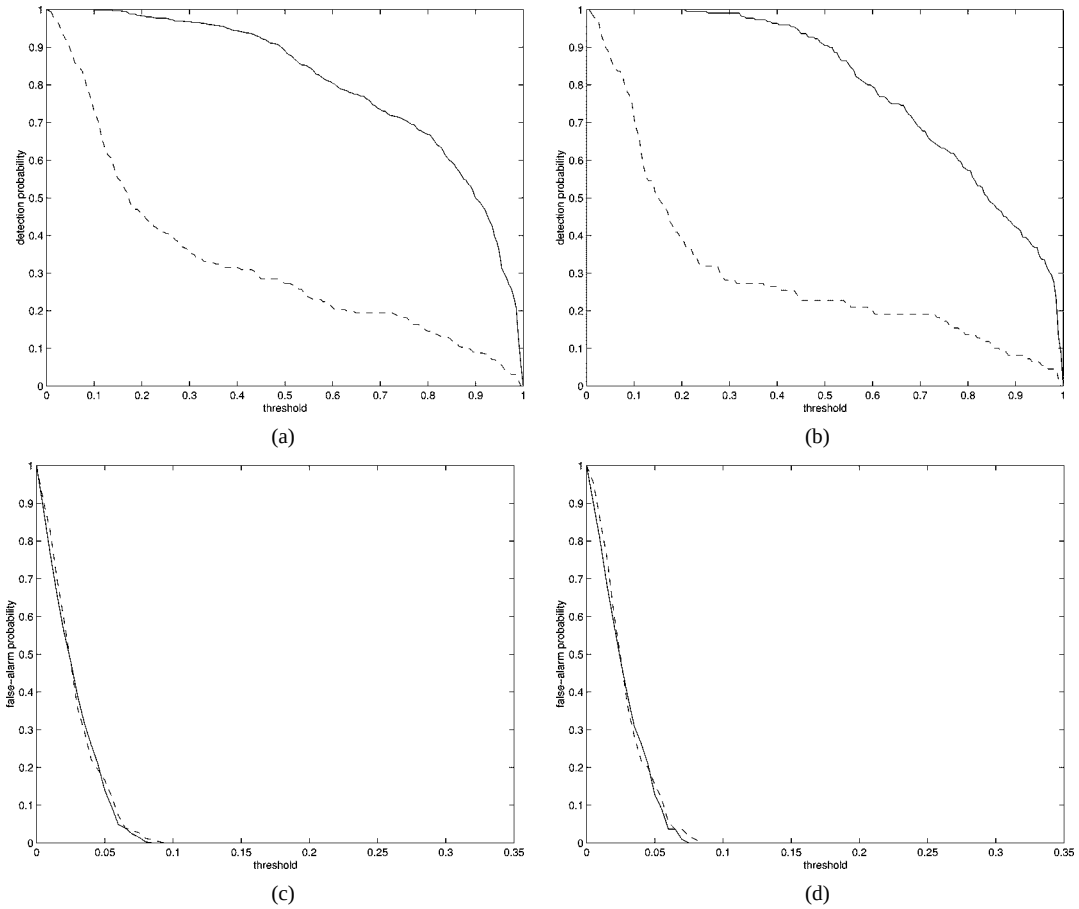


Fig. 5. (a), (b) Mean detection probability and (c), (d) the mean false alarm probability of our method (solid lines) compared with those of Cox's method (dash lines). Left: Attacks 1 to 5 are included. Right: Attacks 1 to 5 are excluded. The horizontal axes of these figures are thresholds. The false alarm probabilities are approximately the same for both methods. Given a threshold, our method has a higher mean detection probability.

cluding both detection with a reference and blind detection, are more effective since they have better average detection values and similar standard deviations than that of Cox *et al.*. There is only a slight difference in the mean and standard deviation between our two detection methods. This can be explained by measuring the mean obtained by projecting the features of each host image onto its watermark space used for detection with a reference. This mean is 0.05, which means that on average, the feature of each host image is approximately perpendicular to the image's watermark space used for detection with the host required as reference.

In Fig. 5, the mean detection probability and the mean false alarm probability of different methods are compared. Our experiments were carried out by applying different attacks to our images and testing the resultant detection values against a threshold. We determined that an image had our watermark if the detected value was higher than the threshold. The detection probability measures the probability that an attacked watermarked image has a detection value greater than a given threshold. The top part of the figure plots the mean detection probability versus a threshold. The mean values are averages of the values of various attacks on our 22 watermarked images. In Fig. 5(a), the previously mentioned 15 attacks applied to these watermarked images are included while in the top right subfigure, only the Attacks 6 to 15 are included. The curves show similar profiles, and ours has higher detection values. The bottom part of the figure plots the mean false alarm probability. The false-alarm probability measures the probability that an image contains no watermark sequence but is falsely identified as ours since its detected value is greater than a given threshold.

Fig. 6 shows the results of the recovered Bee pattern when the watermarked image was subjected to JPEG compression attacks. The quality value in xv for JPEG was set to 40. Shown at the top is the compressed Lena image. Shown in the middle is the recovered Bee pattern, and shown at the bottom is a two-level image produced by thresholding the the middle pattern at zero. A more sophisticated thresholding method can be found in [17]. The left column lists the results obtained using the original image as a reference, while the right column lists the results obtained using blind detection.

B. Blind Attacks

When we carried out the following blind attack experiment, we assumed that our subjects were naive attackers: they did not know our watermarking method. Five subjects were involved in this experiment. All of them understood image processing methods quite well. They were instructed to attack our watermarked Lena image aggressively, modulated with the Bee logo, but to keep the resultant images as visually acceptable as possible. Each of them carried out from 25 to 47 attacks using popular programs like Photoshop version 6, PhotoImpact version 5,

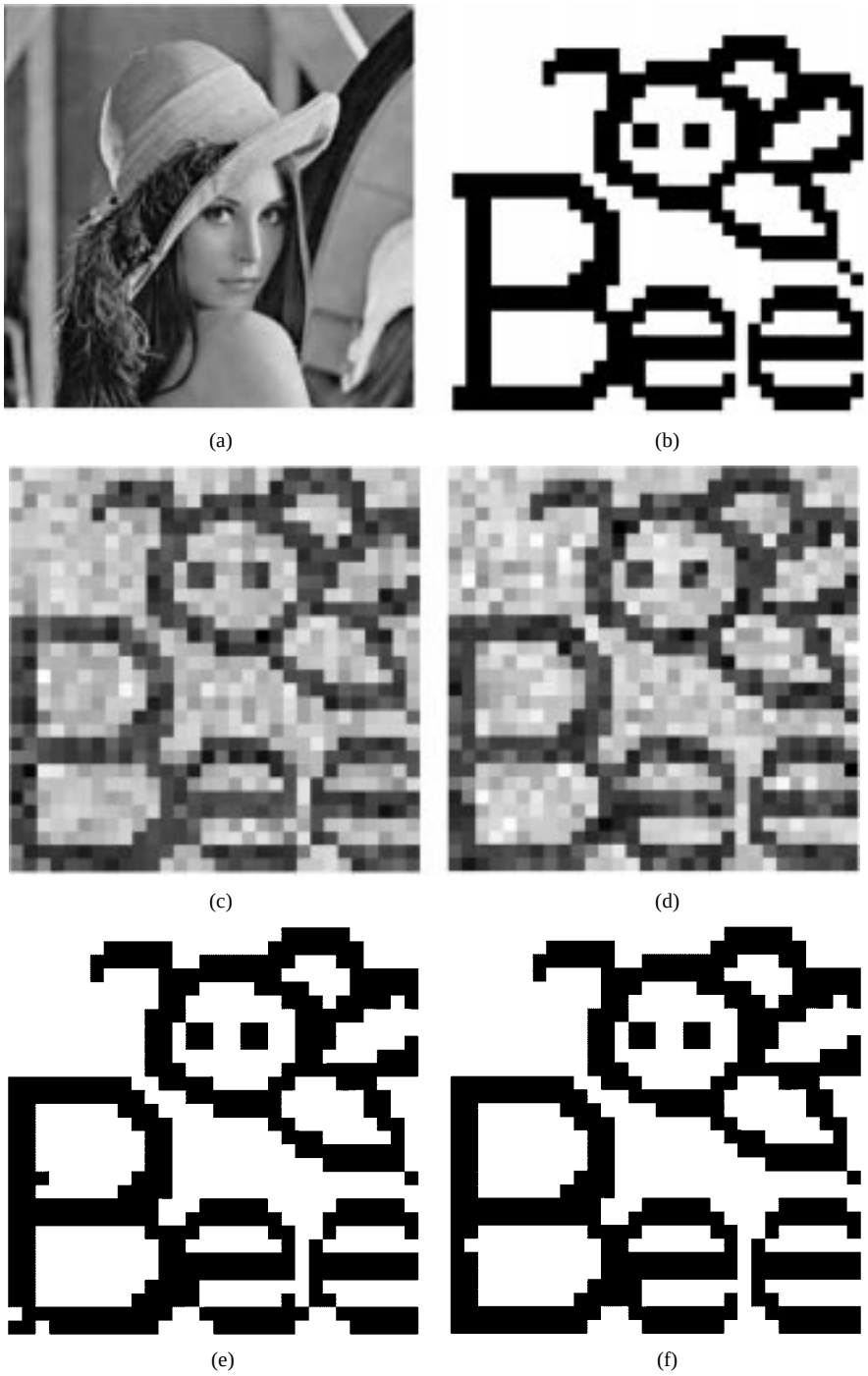
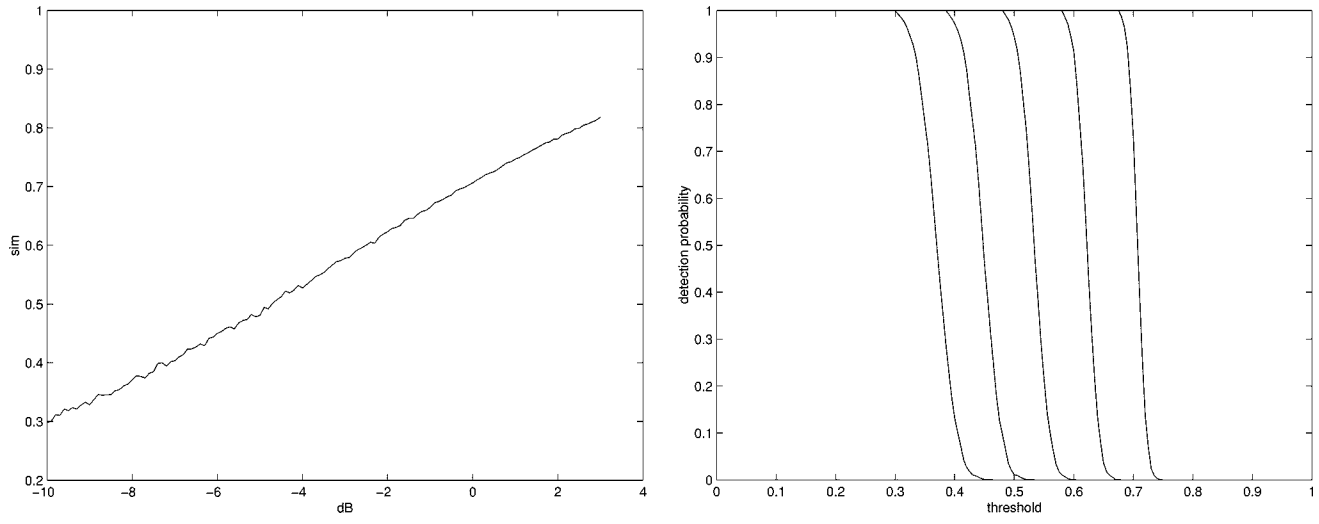


Fig. 6. (Top) JPEG compression with a quality factor of 40. (a) Compressed image and (b) our Bee Pattern. (Middle) Extracted Bee (c) with reference and (d) without reference. (Bottom) Two-level Bee obtained by thresholding the middle image (e) with reference and (f) without reference.

Microsoft Photo Edit version 3, and xv. Most of their attacks were performed by means of elementary operations provided by the programs or combinations thereof. A total of 120 images of the attacked images were selected by means of a voting procedure by our subjects. Our subjects voted for an attacked image if the image was perceptually acceptable. Fig. 8 shows the histogram of the sim value for detection applied to these 120 images. In our experiment, the Lena image was used as the reference image. From the histogram, 86.55% of the attacked images had a detected sim value > 0.5, and about 80.67% of them had a sim value larger than 0.7. So far, we have assumed that the attackers do not know our watermark space. In the Appendix , we present an experiment in which attackers aimed to discover our watermark space and its bases using these 120 images. Our experimental results indicate that our watermark space can be approximately estimated but that its bases can not.

C. Malicious Attacks

The following malicious attacks were carried out by attackers who knew both our watermark spaces, its bases, and our watermarking methods.



(a) (b)

Fig. 7. Spreading random noise attack on W. The watermark space of each image was attacked by random noise 64 times. There were 22 images. (a) Average sim versus SNR. (b) Mean detection probability versus a threshold when attacks were performed on watermarked images. From left to right, the SNR of each curve is, respectively, to be -8, -6, -4, -2, and 0 dB.

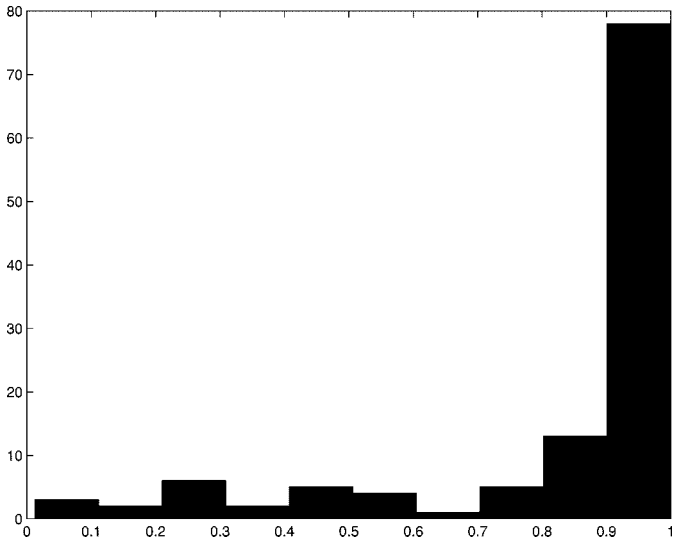


Fig. 8. Histogram of the sim values of 120 test images.

1) Spreading Noise Throughout Watermark Space: In this attack, we assumed that the attackers attacked our watermark space W by spreading random noise in it, hoping that the noise would flood the space and, as a result, remove our watermark feature. We embedded 64 random noises with various levels of energy into the watermark spaces of the previously mentioned 22 images. The performance results for response to this attack are shown in Fig. 7(a). In this figure, we plot the mean, obtained by averaging the detection values of 64 attacks on 22 images, versus SNR, measured by $20\log_{10}||\mathbf{m}^*||/||\mathbf{w}||$, where $\mathbf{m}^*$ and $\mathbf{w}$ are, respectively, to be our watermark and noise feature. In the right part of the figure, we plot the curves of the mean detection probability versus a threshold when SNR from the left to right curves are, respectively, to be -8, -6, -4, -2, and 0 dB.

The perceptual quality of the attacked images were evaluated as follows:

a) Visible distortion from the original image: The visible distortion of an attacked image is measured when a subject can compare the attacked image with the original image. We used a test group of ten persons. All of them knew little about image processing. We asked them to comment on the perceptual quality of each attacked image when it was placed side by side with the reference. When the *SNR* was above -2 dB, none of them saw any difference between the attacked image and the original image. When the SNR was below -4 dB, all the subjects noticed differences between the images, and the average sim at this *SNR* was below 0.5. This implies that if the jamming attack is so severe that the sim falls to below 0.5, according to this test, the attacked image will have noticeable visual distortion compared to the original image.

b) Objectionable distortion: In real-world scenarios, one may not always be able to compare an attacked image with a reference image. Therefore, we asked our subjects (the same ones used in the visual distortion) to comment on whether the perceptual quality of an image was acceptable or not by giving a score for each image. Therefore, the set of attacked images (the same set used in the visual distortion), the watermarked images, and the original images were presented to our subjects one image at a time. The subjects were asked to give a score according to the perceptual quality of each image, but they did not know which image was which. The scores ranged from one to five. A higher score meant better quality. We also asked our subjects to give a score less than three if an image was not acceptable. The results were as follows: the score for -8 dB was 2.1, those for -6 dB, -4 dB, -2 dB were, respectively, 3, 3.6, 3.4, and those for 4 dB, the watermarked images, and the original images were higher than four. According to this experiment, it was not until SNR is below -8 dB, where the mean detection value was below 0.35, that the attacked images were not accepted by our subjects. We conclude that, according to our experiments with these two perceptual distortions, our subjects started to notice visual differences between attacked images and original images when SNR was -4 dB, which means detection value was 0.5. Furthermore, when the *SNR* was -8 dB, with detection value of 0.35, our subjects started to regard the attacked images as perceptually not acceptable.

2) Copy Attacks: Finally, we experimented on our watermarking method using the copy attack [14]. The watermark copy attack attempts to extract an approximation of the watermark from a watermarked (stego) image. Once an approximation of the watermark is obtained, it can be embedded into other (target) images. In the copy attack, the watermark is considered to be noise added to the original image. The watermark prediction process is, therefore, a denoising process. Once an approximation of the watermark is predicted from the stego, one can modify the energy of this predicted watermark and then embed it into a target image. In the experiment, we assumed that the copy attacks were conducted by attackers who knew perfectly our watermarking methods and watermark spaces. Our experiment was carried out as follows.

1) Alice has her watermark signature $\mathbf{m}^*$ of image $\mathbf{Y}$ and Bob has images $\mathbf{Y}$ and $\mathbf{Z}$. Bob knows W_Y and W_Z, the watermark spaces of $\mathbf{Y}$ and $\mathbf{Z}$, respectively.
2) Bob applies the copy attack to the image $\mathbf{Y}$ by
   - estimating the watermark image N using a denoising method;
   - projecting N onto W_Y and obtaining the vector N_Y;
   - increasing the energy of N_Y and then projecting the result onto W_Z. This corresponds to embedding into $\mathbf{Z}$ what Bob believes to be Alice's watermark signature for $\mathbf{Y}$. In this step, the energy of N_Y should be as large as possible but not large enough to distort the visual quality of the resultant image.
3) Alice uses W_Z and her $\mathbf{m}^*$ to extract the watermark signature embedded in $\mathbf{Z}$ by Bob.

We estimated the watermark image N by measuring the local mean of the image $\mathbf{Y}$ with a 3×3 mask. We then carried out the above copy attack on the set of 22 images described in Section IV-A by randomly pairing any two images and denoting any one image in a pair as $\mathbf{Y}$ and the other as $\mathbf{Z}$. Thus, each image belonged to two pairs, and there was a total of two detection values. Their mean and standard deviation were, respectively, 0.0111 and 0.0340, indicating that our watermarking method is robust to copy attacks even when the attackers are informed of our watermarking methods and watermark spaces.

V. Conclusion

In watermark detection, the media content where a watermark is embedded is noise. This content is not viewed purely as noise in our approach. From a set of estimated forged images of the original image, we derive a watermark subspace from the robust feature of the image by means of the second order statistics. A watermark sequence is then embedded into the watermark space using a spreading spectrum. Our watermarking methods are applicable to watermark detection whether a reference image is given or not. The proposed methods can be used to enhance many existing watermarking methods. It may also be possible to find a watermark space using higher order statistics.

Appendix
Attack That Aims to Estimate the Watermark Space

We give an example here of an attack that aims to derive our watermark space. We assume that the attackers know the dimension of our watermark space, but that they do not know the details of our simulations. Thus, he/she had to make up his/her own simulations to obtain his/her watermark space. We use the 120 test images in Section IV-B as the simulation data for watermark space estimation.

Let $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_N]$ and $\mathbf{U}^a = [\mathbf{u}_1^a, \ldots, \mathbf{u}_N^a]$ be our and the attacker's eigenvectors, respectively. Our watermark space, W, and the attacker's, W^a, are the subspaces spanned by the last $N-d$ column vectors $[\mathbf{u}_{d+1}, \ldots, \mathbf{u}_N]$ and $[\mathbf{u}_{d+1}^a, \ldots, \mathbf{u}_N^a]$, respectively. In our experiment, $N - d$ is 900. We use the sequences $\{z_i\}$ and $\{h_i\}$ to find the relationship between $\mathbf{U}$ and $\mathbf{U}^a$.

We define

$$h_i = \sum_{j=d+1}^{N} \left(\mathbf{u}_i^T \, \mathbf{u}_j^a\right)^2$$

where h_i is the energy of our basis $\mathbf{u}_i \in W$ projected onto W^a. The sequence $\{h_i\}_{d+1}^N$, thus, represents the amount of overlap between our watermark space and the attack's watermark space. The mean of $\{h_i\}_{d+1}^N$ is 0.9. Since the value of $\{h_i\}_{d+1}^N$ is relatively large, we conclude that W and W^a have a large area of overlap.

We then define

$$z_i = \max_{j=d+1}^{N} \{|\mathbf{u}_i^T \, \mathbf{u}_j^a|\}$$

for $i = d+1, \ldots, N$. The value z_i is the maximal energy of our basis in W projected onto the attacker's bases in W^a. If z_i is small, then the energy of our basis $\mathbf{u}_i$ is spread out in every direction of the attacker's bases in W^a. Thus, the sequence $\{z_i\}_{d+1}^N$ is a measurement of the alignment of a basis in W with some basis in W^a. We measure that the mean of $\{z_i\}_{d+1}^N$, is 0.1123. Thus, there is only limited alignment of the bases in W and W^a.

This experiment confirms our assumption that most of the attacks on an image have similar main directions in the feature space. The watermark space can be approximately estimated, while its bases are harder to estimate.

Acknowledgment

This manuscript was significantly improved by all the reviewers. Special thanks go to the first reviewer, who pointed out that the principle behind our method may be related to the principle proposed in [8] where it is advocated that content not be purely viewed as noise in watermark detection.

References

[1] L. Boney, A. H. Tewfik, and K. N. Hamdy, "Digital watermarks for audio signals," in *Proc. Multimedia 1996*, 1996, pp. 473–480.
[2] W. Bender, D. Gruhl, and N. Morimoto, "Techniques for data hiding," *Proc. SPIE*, pp. 452–455, June 1995.
[3] G. W. Braudaway, K. A. Magerlein, and F. Mintzer, "Protecting publicly-available images with a visible watermark," *Proc. SPIE*, vol. 2659, pp. 126–133, 1996.
[4] C. H. Chou and Y. C. Li, "A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 5, pp. 467–476, Dec. 1995.
[5] I. J. Cox, J. Kilian, T. Leighton, and T. Shamoon, "A secure, robust watermark for multimedia," in *Proc. Workshop on Information Hiding*, May 1996.

[6] ——, "Secure spread spectrum watermarking for multimedia," *IEEE Trans. Image Processing*, vol. 6, pp. 1673–1687, Dec. 1997.
[7] I. J. Cox and J. P. Linnartz, "Some general methods for tampering with watermarks," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 587–593, May 1998.
[8] I. J. Cox, M. L. Miller, and A. L. Mckellips, "Watermarking as communications with side information," *Proc. IEEE*, vol. 87, pp. 1127–1141, July 1999.
[9] S. Craver, N. Memon, B. L. Yeo, and M. M. Yeung, "Resolving rightful ownerships with invisible watermarking techniques: Limitations, attacks, and implications," *IEEE J. Select. Areas Commun.*, pp. 573–586, May 1998.
[10] F. Hartung and M. Kutter, "Multimedia watermarking techniques," *Proc. IEEE*, vol. 87, July 1999.
[11] I. Hontsch, L. J. Karam, and R. J. Safranek, "A perceptually tuned embedded zerotree image coder," in *Proc. IEEE Conf. Acoustics, Speech, Signal Processing*, 1989, pp. 1945–1948.
[12] C. T. Hsu and J. L. Wu, "Hidden digital watermarks in images," *IEEE Trans. Image Processing*, vol. 8, pp. 58–68, Jan. 1999.
[13] D. Kundur and D. Hatzinakos, "Digital watermarking using multiresolution wavelet decomposition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 5, 1998, pp. 2969–2972.
[14] M. Kutter, S. Voloshynovskiy, and A. Herrigel, "The watermark copy attack," *Proc. SPIE*, vol. 3971, pp. 371–380, 2000.
[15] P. Moulin and J. A. O'Sullivan, "Information-theoretic analysis of watermarking," in *Proc. IEEE ICASSP 2000*, Istanbul, Turkey, June 2000.
[16] P. Moulin and E. Delp, "A mathematical approach to watermarking and data hiding," in *Proc. IEEE ICIP 2001*, Thessaloniki, Greece, Oct. 2001.
[17] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, Jan. 1979.
[18] F. A. P. Petitcolas, R. J. Anderson, and M. G. Kuhn, "Attacks on copyright marking systems," *Lecture Notes Comput. Sci. 1525: Information Hiding*, pp. 218–238, 1998.
[19] ——, "Information hiding–A survey," *Proc. IEEE*, pp. 1062–1078, 1999.
[20] H. V. Poor, *An Introduction to Signal Detection and Estimation*. New York: Springer, 1994.
[21] I. Pitas, "A method for watermark casting on digital images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, pp. 775–780, Oct. 1998.
[22] M. D. Swanson, M. Kobayashi, and A. H. Tewfik, "Multimedia data-embedding and watermarking technologies," *Proc. IEEE*, vol. 86, pp. 1064–1087, June 1998.
[23] M. M. Yeung, "Digital watermarking," *Commun. ACM*, vol. 41, no. 7, pp. 31–33, July 1998.
[24] W. Zeng and B. Liu, "A statistical watermark detection technique without using original images for resolving rightful ownerships of digital images," *IEEE Trans. Image Processing*, vol. 8, pp. 1534–1548, Nov. 1999.

Jengnan Tzeng received the B.S. degree from National Chengchi University, Taipei, Taiwan, R.O.C., in 1995 and the M.S. degree from the National Central University, Taoyuan, Taiwan, in 1997, where he is currently pursuing the Ph.D. degree student in mathematics.

His current research interests include digital watermarking method, wavelets analysis, wavelets, and PDE methods in image processing and computer vision.

Wen-Liang Hwang received the B.S. degree in nuclear engineering from National Tsing Hua University, Hsinchu, Taiwan, R.O.C., the M.S. degree in electrical engineering from the Polytechnic Institute of New York, Brooklyn, and the Ph.D. degree in computer science from New York University in 1993.

He was a Postdoctoral Researcher with the Department of Mathematics, University of California, Irvine, in 1994. He became a Member of the Institute of Information Science, Academia Sinica, Taipei, Taiwan, in January 1995. He is currently as an Associate Research Fellow. His research interests include wavelet analysis, signal and image processing, multimedia communication, and computer vision.

I-Liang Chern received the B.S. and M.S. degrees in mathematics from National Taiwan University, Taipei, and the Ph.D. degree in mathematics from New York University in 1983.

He has been with Academia Sinica, Taipei, the Courant Institute, New York University, and the Argonne National Laboratory, Argonne, IL. He has been a Professor with the Mathematics Department, National Taiwan University, since 1991. His research interests include multiscale scientific computing, wavelet analysis, and partial differential equations.